УДК 004.428.2

Є. Є. Герасімов
І. В. Богач

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТРИК ВІДСТАНІ ДЛЯ ВЕКТОРНОГО ПОШУКУ

Вінницький національний технічний університет

Анотація

В роботі проведено порівняльний аналіз метрик відстані для векторного пошуку. Оцінено евклідову відстань, косинусну подібність, манхеттенську відстань, подібність Жаккарда та відстань Махаланобіса, досліджено їхній вплив на точність пошуку, ефективність обчислення.

Ключові слова: Векторний пошук, метрика відстані, обробка природної мови (NLP), Евклідова відстань, косинус-подібність, Манхеттенська відстань, подібність Жаккарда, відстань Махаланобіса.

Abstract

In the paper, a comparative analysis of distance metrics for vector search is carried out. Euclidean distance, cosine similarity, Manhattan distance, Jaccard similarity, and Mahalanobis distance were evaluated, and their influence on search accuracy and calculation efficiency was investigated.

Keywords: Vector search, distance metrics, Natural Language Processing (NLP), Euclidean distance, cosine similarity, Manhattan distance, Jaccard similarity, Mahalanobis distance.

Вступ

В останні роки в галузі Natural Language Processing (NLP) спостерігається сплеск використання векторних репрезентацій тексту для різних завдань, таких як пошук інформації, класифікація документів і оцінка семантичної подібності. Векторний пошук відіграє вирішальну роль у цих програмах, забезпечуючи ефективний пошук відповідної інформації з великих наборів даних. Одним із ключових компонентів, який суттєво впливає на продуктивність векторного пошуку, є вибір метрики відстані, яка використовується для вимірювання подібності між векторами.

Теза спрямована на проведення комплексного порівняльного аналізу метрик відстані, які зазвичай використовуються у векторному пошуку в контексті додатків NLP. Дослідження зосереджено на оцінці ефективності цих показників з точки зору точності пошуку, обчислювальної ефективності та стійкості до шуму та розмірності.

Результати дослідження

Дослідження оцінювало такі показники відстані:

1. Евклідова відстань: ця метрика обчислює відстань по прямій лінії між двома векторами в евклідовому просторі. Це широко використовувана метрика, але вона не може ефективно вловити семантичну подібність.
2. Косинусна подібность: косинус подібності вимірює косинус кута між двома векторами, забезпечуючи міру їх подібності незалежно від їх величини. Він часто використовується в завданнях NLP через його стійкість до довжини вектора.
3. Манхеттенська відстань: також відома як відстань L1, манхеттенська відстань обчислює суму абсолютних різниць між відповідними елементами двох векторів. Він підходить для випадків, коли велика розмірність.
4. Подібність за Жаккардом: подібність за Жаккардом вимірює перетин над об’єднанням ненульових вимірів у двох двійкових векторах. Він зазвичай використовується для розріджених векторів у аналізі тексту.
5. Відстань Махаланобіса: цей показник враховує структуру коваріації даних і особливо корисний під час роботи з корельованими функціями.

Picture	Method	Application	Features	Disadvantages	Formula
	**Euclidean Distance**	General distance measurement, Clustering, Classification, Regression	Measures the straight line distance between two points in n-dimensional space.	Sensitive to outliers, Can be affected by scale differences	O(n) Fast
	**Manhattan Distance**	Distance on grid networks, Routing algorithms, Image processing	Measures the distance between two points on a grid network, where movement is limited.	Ignores diagonal movement, not useful for high-dimensional data,	O(n) Fast
	**Cosine Similarity**	Text document clustering, Text analysis, Recommendation systems	Measures the cosine of the angle between two vectors	Ignores magnitude of vectors, Not useful for negative values or high degree of correlation data	O(n) Fast
	**Jaccard Similarity**	Set similarity measurement, Text analysis, recommendation systems	Measures the similarity between two sets by comparing their intersection and union.	Ignores magnitude of sets, May not be as useful for continuous data	O(n) Fast
	**Mahalanobis Distance**	Multivariate statistical analysis, Outlier detection, Clustering	Measures the distance between two points in n-dimensional space, taking into account the correlation between variables.	Requires full covariance matrix, May not be as useful for datasets with a large number of variables	O(n^3) Slow

Рисунок 1 – Порівняння метрик відстані

Оцінка проводилася на еталонних наборах даних для таких завдань, як пошук документів, оцінка семантичної подібності та кластеризація. Результати показують, що вибір метрики відстані значно впливає на продуктивність алгоритмів векторного пошуку. Косинусна подібність незмінно переважала інші показники з точки зору точності пошуку та стійкості до шуму. Евклідова відстань продемонструвала конкурентоспроможність, але мала проблеми з великомірними даними. Манхеттенська відстань і подібність Жаккарда продемонстрували хорошу ефективність у конкретних сценаріях, таких як розріджені дані та висока розмірність відповідно. Відстань Махаланобіса, хоч і ефективна для виявлення кореляції ознак, показала вищі обчислювальні витрати.

Висновок

На завершення порівняльний аналіз підкреслює важливість вибору відповідної метрики відстані на основі конкретних вимог завдання NLP. Косинусна подібність стає надійним вибором для векторного пошуку в більшості сценаріїв, пропонуючи баланс між точністю та обчислювальною ефективністю. Однак для особливих випадків, таких як розріджені дані або корельовані функції, інші показники, такі як подібність Жаккарда або відстань Махаланобіса, можуть забезпечити кращу продуктивність. Майбутні напрямки досліджень включають вивчення гібридних показників відстані та включення предметно-специфічних знань для покращення можливостей векторного пошуку в програмах NLP.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Vector Similarity Explained [Електронний ресурс] – Режим доступу до ресурсу: https://www.pinecone.io/learn/vector-similarity

2. An Empirical Study on the Performance of the Distance Metrics [Електронний ресурс] – Режим доступу до ресурсу: https://dergipark.org.tr/en/download/article-file/3257351 .

3. An Exhaustive List Of Distance Metrics For Vector Similarity Search [Електронний ресурс] – Режим доступу до ресурсу: https://medium.datadriveninvestor.com/an-exhaustive-list-of-distance-metrics-for-vector-similarity-search-09c4db84f0b4.

4. Exploring Common Distance Measures for Machine Learning and Data Science: A Comparative Analysis [Електронний ресурс] – Режим доступу до ресурсу: https://medium.com/@eskandar.sahel/exploring-common-distance-measures-for-machine-learning-and-data-science-a-comparative-analysis-ea0216c93ba3

Герасімов Євген Євгенович *– студент групи 1АКІТ-20б, кафедра автоматизації та інтелектуальних інформаційних технологій, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м.Вінниця, e-mail: yevhen.gerasimov@gmail.com*

Богач Ілона Віталіївна *– к.т.н., доцент кафедри автоматизації та інтелектуальних інформаційних технологій, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м.Вінниця, e-mail: ilona.bogach@gmail.com*

Herasimov Yevhen Yevhenovych *– student of 1ACIT-20B group, Department of Automation and Intelligent Information Technologies, Faculty of Intelligent Information Technology and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: yevhen.gerasimov@gmail.com*

Bogach Ilona Vitaliivna *– Associate Professor of Automation and Intelligent Information Technologies, Faculty of Computer Systems and Automatics Vinnytsia National Technical University, Vinnytsia, e-mail: ilona.bogach@gmail.com.*